Cogn-IQ Research Papers Validation Study ◆ Open Access doi: pubscience.org/ps-1n0JV-785570-6dGh

WN RELIABILITY

Reliability of the What's Next Number-Sequence Test: Internal Consistency and Jouve's Randomized Reliability Estimate on an Independent Blind-Rescored Online Sample

Xavier Jouve

Head of Research

Cogn-IQ

Author Note

Xavier Jouve, Head of Research, Cogn-IQ. ORCID iD: 0009-0006-1283-045X.

The author is the developer of the What's Next (WN) test evaluated here and of the randomized reliability estimator applied to it. These potential conflicts of interest are disclosed in the interest of full transparency. No external funding was received for this work.

Correspondence concerning this article should be addressed to Xavier Jouve, Cogn-IQ. Email: research@cogn-iq.org

ABSTRACT

The What's Next (WN) is an 81-item fixed-form number-series test of inductive reasoning whose scores are reported as the Logical Inference Index (LII; $M = 100$, $SD = 15$). This study re-establishes the instrument's reliability on an independent sample of 210 protocols harvested from routine online administration between November 2024 and July 2026, scored blind against the production multi-variant key after an attention screen and a first-administration deduplication. Internal consistency was high (KR-20 = .95), and the mean corrected item-rest correlation was .45, with the least discriminating items being the easiest ones. Jouve's Randomized Reliability Estimation, the Monte-Carlo mean of Spearman–Brown-adjusted random split-half reliabilities, gave $\xi = .96$; the 95% interval across arbitrary item partitions was only .03 wide and every random partition cleared the .90 standard, so the coefficient does not depend on how the form is halved. The implied standard error of measurement was about three LII points. The estimates held within each of the two form versions present in the sample (KR-20 = .96 and .93) and reproduced the published calibration figures (KR-20 = .965; $\xi = .97$) on a fresh, independently scored sample more than four times the size of the earlier study. The WN is a highly reliable measure of inductive reasoning.

Keywords: reliability, internal consistency, randomized reliability estimation, inductive reasoning, number series

Introduction

The What's Next test (WN) is an 81-item fixed-form measure of inductive reasoning by number-series completion, reporting a single standard score, the Logical Inference Index (LII; $M = 100$, $SD = 15$). Its reliability has been documented in three earlier studies: the classical internal consistency of the calibration sample was KR-20 = .965 ($N = 435$; Sharma, 2024), a randomized reliability estimate of $\xi = .97$ was reported alongside an item-level scaling study ($n = 47$; Jouve, 2025a), and the earliest characterization gave a coefficient alpha of .93 ($N = 506$; Jouve, 2023). These estimates rest on the calibration and development samples; none has been checked against an independent, later cohort, and the ξ estimate in particular rests on only 47 examinees.

The present study addresses both gaps using a large archive of routine online administrations, scored independently of the operational engine. Two questions organize it. We ask: (RQ1) Does the WN's internal consistency replicate on an independent, recently administered sample scored blind from the raw responses? (RQ2) What value does Jouve's Randomized Reliability Estimation (ξ) take on a sample large enough to characterize the estimator's distribution across the space of item partitions, rather than on a single split?

Method

Sample

Protocols were drawn from routine online administrations of the WN completed between November 2024 and July 2026. After the screens described below, the analysis sample comprised 210 respondents (age $M = 26.3$ years, $SD = 9.1$). Because the WN is a self-selected online instrument, the sample is above average in ability and somewhat restricted in range relative to the general population; this is a property of the testing

population, not of the analysis, and its effect on the coefficients is considered in the Discussion.

Blind Rescoring

Each protocol was scored from its raw item responses rather than from the platform's stored total. The accepted-answer set for every item was parsed directly out of the production scorer so that the key used here stays in lockstep with the operational one, including its multiple-solution rubric: an item is credited when the supplied continuation matches any admissible solution the item's structure permits (Jouve, 2023). The raw number correct across the 81 items served as the total score. No item content or key entry is reproduced here, to protect item security.

Screens

Three screens were applied before analysis. First, only each respondent's first administration was retained, deduplicating on a normalized email-then-fingerprint key so that repeated sittings by the same person could not inflate consistency. Second, developer and author submissions were excluded. Third, an attention screen required that both of two easy anchor items positioned late in the fixed form had been attempted, which indicates that the respondent worked through to the end of the form rather than abandoning it early; unattempted late items are otherwise scored as incorrect, consistent with the operational total.

Analysis

Internal consistency was estimated by the Kuder–Richardson formula 20 for dichotomous items, $\mathrm{KR\text{-}20} = \frac{k}{k-1}\left(1 - \sum_i p_i q_i / \sigma_X^2\right)$, where $k = 81$, p_i is the proportion passing item i, $q_i = 1 - p_i$, and σ_X^2 is the total-score variance (Kuder & Richardson, 1937; Cronbach, 1951). Item quality was summarized by the mean corrected item-rest correlation. Reliability was also estimated by Jouve's Randomized Reliability Estimation (ξ), the estimator introduced by Jouve (2025b) and computed with the canonical calculator (Jouve, 2025c). Rather than reduce reliability to one arbitrary split, ξ samples the space of item partitions by Monte Carlo: at each iteration the items are randomly partitioned into two balanced halves by a Fisher–Yates shuffle, the half-test scores S_A

and S_B are correlated, and that correlation is Spearman–Brown-adjusted, $r' = 2r_{AB}/(1 + r_{AB})$ (Brown, 1910; Spearman, 1910). The estimate is the Monte-Carlo mean of r' over the sampled partitions, $\xi = \mathbb{E}_S[\,2r_{AB}(S)/(1 + r_{AB}(S))\,]$; the iteration count adapts to the matrix size and the run stops on convergence of the running mean, and the 2.5th and 97.5th percentiles of the partition distribution give a 95% interval. Averaging across partitions is what removes the dependence of an ordinary split-half coefficient on a single arbitrary halving. The standard error of measurement on the reporting metric was obtained as $\mathrm{SEM} = \mathrm{SD}\sqrt{1 - \rho}$ with $\mathrm{SD} = 15$ (Feldt & Brennan, 1989).

Results

The 210 retained respondents had a raw score of $M = 24.46$ ($SD = 13.15$) on the 0-to-81 range, closely matching the calibration sample ($M = 23.56$, $SD = 15.41$; Sharma, 2024). Internal consistency was high, KR-20 = .95, and the mean corrected item-rest correlation was .45 (range −.01 to .70). The few items with weak item-rest correlations were the easiest ones, whose near-uniform pass rates leave little variance to correlate with the total; this is the expected behaviour of easy anchor items, not a sign of construct-irrelevant content. Table 1 collects the coefficients.

Table 1
Reliability and precision of the 81-item WN form (N = 210)

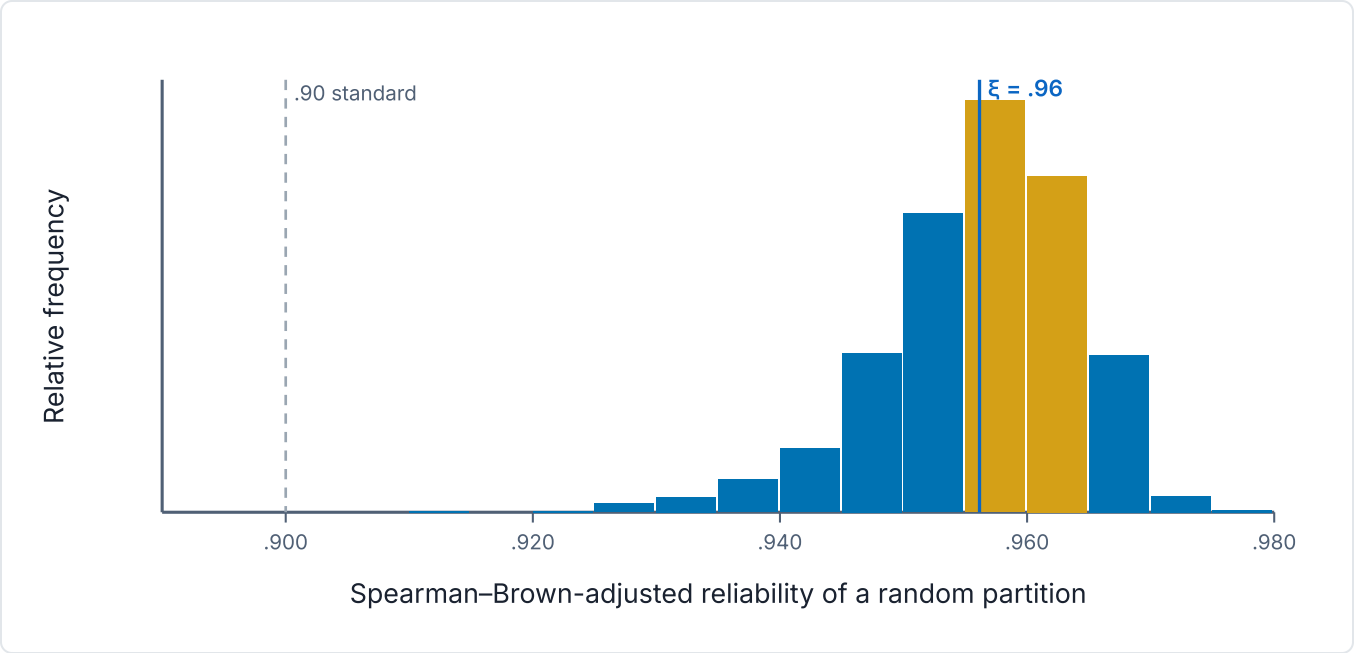
Statistic	Value
KR-20 internal consistency	.95
Mean corrected item-rest correlation	.45 [−.01, .70]
Jouve's ξ (Monte-Carlo mean)	.96
95% partition interval	[.938, .969]
Width of the 95% partition interval	.031
Proportion of partitions reaching .90	1.000
Standard error of measurement (LII points)	3.1

Note. Item-rest range in brackets. ξ is the Monte-Carlo mean of the Spearman–Brown-adjusted correlation over random balanced item partitions; the interval is the 2.5th–97.5th percentile of that partition distribution. The standard error of measurement is computed on the LII metric (*SD* = 15) from ξ.

Jouve's ξ was .96, and its partition distribution was tight: the central 95% interval was [.938, .969] — a width of only .031, well inside the .05 that marks a very stable coefficient — and not one sampled partition fell below the .90 standard. The reliability of the WN therefore does not depend on which arbitrary half of the form is set against the other. Figure 1 shows the partition distribution: it is tightly concentrated near .96 and bounded well above the high-stakes threshold. The implied standard error of measurement is about three LII points, so a 95% confidence interval around an observed LII spans roughly ±6 points.

Figure 1

Partition distribution of the Spearman–Brown-adjusted reliability underlying Jouve's ξ for the WN



Note. Relative frequency of the Spearman–Brown-adjusted reliability over random balanced item partitions (.005-wide bins); the two bins straddling the mean (.955–.960) are highlighted, the solid line marks ξ = .96, and the dashed line marks the .90 high-stakes standard, which every partition clears.

The estimates were stable across the two form versions present in the archive. Rescoring and analyzing each version separately (Table 2) gave KR-20 = .96 and ξ = .96 for the earlier version ($n = 144$) and KR-20 = .93 and ξ = .94 for the later version ($n = 66$); the

slightly lower later-version figures track its narrower score spread ($SD = 10.6$ vs. 14.2) and fewer examinees (66 vs. 144), not a difference in scoring. Pooling the versions is therefore warranted, and the pooled estimates are not an artifact of mixing them.

Table 2
Reliability within each form version and pooled

Sample	*n*	KR-20	ξ	Raw *SD*
Earlier form version	144	.96	.96	14.2
Later form version	66	.93	.94	10.6
Pooled	210	.95	.96	13.2

Note. All protocols were scored blind against the current multiple-solution key. ξ = Jouve's Randomized Reliability Estimate (Monte-Carlo mean). The lower later-version coefficients reflect that version's narrower score variance and smaller sample.

Discussion

Both questions are answered affirmatively. On RQ1, internal consistency replicated closely on an independent, later, blind-rescored sample: KR-20 = .95 here against .965 in the calibration study and .93 in the earliest characterization (Sharma, 2024; Jouve, 2023). That a fresh cohort, scored independently of the operational engine, reproduces the published figure is the strongest available evidence that both the coefficient and the scoring are sound. On RQ2, Jouve's ξ was .96 on 210 respondents — more than four times the sample of the earlier estimate ($n = 47$; Jouve, 2025a) — and its partition distribution was narrow (95% width .031) with every sampled partition clearing the .90 standard, so the coefficient is not contingent on any particular halving of the form (Jouve, 2025b). The implied SEM of about three LII points quantifies the precision a user should attach to an individual score.

Two limitations bound these conclusions. The sample is self-selected and above average, so its score variance is somewhat restricted relative to the general population; because reliability coefficients depend on variance, the estimates are best read as lower bounds for a less restricted population rather than as figures that would shrink in one. And a

small number of easy items contribute little to internal consistency, as their item-rest correlations show; they anchor the low end of the difficulty range and serve the attention screen rather than adding discrimination, and no item showed the negative discrimination that would indicate a mis-keyed or construct-irrelevant item (Nunnally & Bernstein, 1994). Neither limitation qualifies the central result.

Conclusion

On an independent sample of 210 online administrations scored blind against the production key, the What's Next test reproduced its published reliability: internal consistency of .95 and Jouve's ξ of .96, whose narrow partition distribution and universal clearance of the .90 standard establish that the coefficient is not an artifact of a single split. The WN is a highly reliable measure of inductive reasoning, and the LII should be reported as an estimate carrying a standard error of about three points.

Figures & tables

Figure 1. *Partition distribution of the Spearman–Brown-adjusted reliability underlying Jouve's ξ for the WN.*

Table 1. *Reliability and precision of the 81-item WN form (N = 210).*

Table 2. *Reliability within each form version and pooled.*

References

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). American Council on Education & Macmillan.

Jouve, X. (2023). *What's Next IQ test: Psychometric properties*. Cogn-IQ Research Papers. https://pubscience.org/ps-1mSMq-1126f1-fqxh

Jouve, X. (2025a). *Numbers as symbols: An item-level multidimensional scaling of the What's Next number-sequence test and the JCCES*. Cogn-IQ Research Papers. https://pubscience.org/ps-1mYgv-e7bac6-1l7r

Jouve, X. (2025b). *When alpha fails: Jouve's randomized reliability estimation versus classical reliability coefficients*. Cogn-IQ Research Papers. https://pubscience.org/ps-1mYdi-014f7f-ormI

Jouve, X. (2025c). *Jouve's Randomized Reliability Estimation (ξ) calculator* [Computer software]. Cogn-IQ. https://www.cogn-iq.org/statistical-tools/jouve-randomized-reliability-estimation/

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Sharma, P. (2024). *What's Next? (WN) test: 2PL calibration and IQ norms linked to JCCES and WAIS-IV reference scales*. Cogn-IQ Research Papers. https://pubscience.org/ps-1n0JG-7ba35a-Se60

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

CITE THIS ARTICLE

Jouve, X. (2026). Reliability of the What's Next number-sequence test: Internal consistency and Jouve's randomized reliability estimate on an independent blind-rescored online sample. *Cogn-IQ Research Papers.* https://pubscience.org/ps-1n0JV-785570-6dGh

Export: BibTeX · RIS · EndNote · APA · Plain text

OUTLINE

Abstract
Introduction
Method
Results
Discussion
Conclusion
References

ARTICLE

JOURNAL
Cogn-IQ Research Papers

TYPE
Reliability Study · Validation

PUBLISHED

26 Jul 2026

ACCESS
Open Access

DOI
pubscience.org/ps-1n0JV-785570-6dGh

TOOLS

- Full text
- Figures & tables
- Cite this article
- References

CONTINUE ON COGN-IQ

- What's Next (WN) Test — The number-sequence reasoning test whose reliability is evaluated here
- Numbers as Symbols: Item-Level MDS of the WN — The item-level structure and the earlier randomized reliability estimate (ξ)
- Internal Consistency — The reliability concept indexed by KR-20 and the randomized split-half estimator